DOCUMENT RESUME

ED 151 730 CS 003 954

AUTHOR Henry, G.
TITLE The Relation between Linguistic Factors of Written Style Identified by a Principal Components Analysis and Reading Comprehension as Measured by Cloze Tests.
PUB DATE [77]
NOTE 12p.; Study prepared at University of Liege, Belgium

EDRS PRICE MF-$0.83 HC-$1.67 Plus Postage.
DESCRIPTORS *Cloze Procedure; Foreign Countries; *Linguistic Difficulty (Inherent); *Readability; *Reading Comprehension; *Reading Research; Secondary Education; Semantics; Syntax; Written Language
IDENTIFIERS *Belgium

ABSTRACT

Some 2700 Belgian students (grades five through twelve) participated in this study of the relationships between linguistic features of passages and passage readability estimated from cloze procedures. Clozed versions of each of 60 test passages were prepared and administered to students at various grade levels. Twenty-five linguistic features of written style were also identified and scored for each passage. Factor analysis of the 25 linguistic scores yielded a content factor (didactic and descriptive reading versus light reading), a syntactical complexity factor, and a redundancy factor. The three factors together, however, accounted for only 55% of the total variance. Regression analyses indicated that the linguistic features were of varying importance to comprehension at different ages. Semantic content seemed to pose the main difficulty for the youngest children; older children were more able to use additional types of clues (such as repetitions of words) and were more sensitive to syntactic difficulties. (AA)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED151730

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY.

THE RELATION BETWEEN LINGUISTIC FACTORS OF WRITTEN STYLE IDENTIFIED BY A PRINCIPAL COMPONENTS ANALYSIS AND READING COMPREHENSION AS MEASURED BY CLOZE TESTS

G. HENRY

University of Liège

(Belgium)

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

G. Henry

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) AND USERS OF THE ERIC SYSTEM."

The literature is replete with studies concerned with the construction of readability formulas, but contains few studies relevant to the resolution of theoretical problems related to readability, despite the pertinence of many of these problems to an understanding of the reading comprehension process. In addition, most of the studies of theoretical interest are limited in generalisability because of either a restricted range of variables or because of a failure to extend the study to include an analysis of the relationships between factor scores and comprehension criterions.

Brinton and Danielson (1958), Stolurow and Newman (1959) conducted some important studies in this domain. Bormuth's study (1969) was highly sophisticated but this author did not compare the relationships existing between linguistic variables and comprehension criterions at different grade levels.

The purpose of the present study was therefore : to factor analyse a large number of linguistic features of written style; to relate the factors which could be identified to the difficulty of comprehension of passages; and to compare the influence of these factors at different grade levels.

C 5003 954

PROCEDURE

Identification of linguistic features

A review of the literature produced twenty-five linguistic features of written style which could be objectively measured and which were appropriate to the content of school textbooks. Each reading passage was measured on each linguistic feature. The twenty-five resulting variables scores for each reading passage provided the data for a principal components analysis (with varimax rotation). The variables used are listed in table 1.

Table 1

LINGUISTIC VARIABLES

1	Mean sentence length (number of words)
2	Mean word length (number of letters)
3	Type-token ratio
4	Word frequency count (absentees from a basic vocabulary list) (CREDIF list) (Gougenheim, et al., 1967)
5	Suffixed nouns
6	Proper nouns (including adjectivials)
7	Concrete nouns
8	Personal pronouns (3rd person)
9	Personal pronouns (1st and 2nd person)
10	Articles and deictics
11	Ratio of adjectives before nouns to adjectives after nouns
12	Adjectives (in the noun phrase employed as attributes)
13	Ratio of animate subjects to inanimate subjects
14	Short adverbs + adverbials + presentatives
15	Prepositions
16	Conjunctions
17	Passives
18	Comparatives and superlatives
19	Infinitives
20	Complement and subject sentences
21	Reciprocals and reflexives
22	Shift transformations
23	Gerunds and participles
24	Dialog variable block (including addresses + exclamation marks and quotation marks)
25	Relative clauses and subordinate sentences (except : time clauses)

Construction of cloze tests

A cloze readability test is made by deleting every fifth word from a passage and replacing the deleted words with underlined blanks of a uniform length. For each passage, five forms were made by deleting words 1, 6, 11, ... to make the first form, 2, 7, 12, ... to make the second form, etc.

Data Collection

About 2700 students were tested. All were enrolled in public school systems in the suburbs of Liège and Bruxelles (Belgium). Approximately 900 of the subjects were enrolled in grades 5 and 6 (level I), 900 in grades 8 and 9 (level II) and 900 in grades 11 and 12 (level III).

Every cloze test form (1-5) was then randomly assigned to 12 students at the three school levels, and the passage difficulty of each passage was calculated by averaging word difficulties across the passage.

The reliability of the cloze test scores was estimated for each school level (I to III) by dividing each 60 subjects sample into two random of 30 subjects, calculating two estimates of the passage's difficulty and correlating across all passages the two sets of means. These correlations ranged from .95 to .97 for the three school levels.

RESULTS

Factor Analysis

The intercorrelation matrix calculated for all of the 25 variables is shown in Table 2. This correlation matrix among the 25 variables was then factor analysed, using a principal components method and an orthogonal rotation of the factor matrix. Three factors with eigenvalues greater than 1.0 were found. The factor loadings of the 25 variables on these three factors are given in Table 3.

Table 2

CORRELATION MATRIX BETWEEN LINGUISTIC VARIABLES

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
1	1.00	.55	.06	.44	.53	.52	-.69	-.59	-.38	.37	-.56	-.24	-.54	-.50	.47	.39	.58	.40	.52	.19	.30	.59	.59	-.53	.65
2		1.00	.09	.67	.66	.47	-.57	-.49	-.50	.62	-.59	-.30	-.69	-.48	.38	.11	.50	.22	.00	.09	.09	.42	.28	-.73	.28
3			1.00	.22	.17	.30	-.10	-.16	-.06	-.15	.00	.04	.05	.28	-.01	-.08	-.10	.04	.22	.23	-.08	.17	.01	.00	-.01
4				1.00	.64	.37	-.35	-.37	-.58	.64	-.61	-.21	-.68	-.46	.35	.16	.35	.15	.13	.08	.14	.34	.37	-.74	.21
5					1.00	.55	-.68	-.58	-.44	.40	-.53	-.33	-.64	-.31	.46	.14	.57	.29	.27	.23	.01	.45	.34	-.66	.30
6						1.00	-.47	-.53	-.39	.26	-.13	-.20	-.37	-.23	.52	.01	.48	.11	.05	-.15	-.03	.40	.35	-.39	.16
7							1.00	.57	.35	-.34	.46	.24	.51	.40	-.43	-.29	-.56	-.35	-.42	-.25	-.08	-.54	-.37	.62	-.48
8								1.00	.23	-.37	.55	.33	.57	.16	-.46	-.23	-.53	-.23	-.16	-.05	.13	-.37	-.46	.59	-.36
9									1.00	-.68	.24	.12	.41	.28	-.41	-.18	-.31	-.08	-.07	.08	.11	-.45	-.34	.58	-.10
10										1.00	-.43	-.34	-.62	-.54	.31	.26	.51	.16	-.01	-.05	-.16	.36	.35	-.64	.17
11											1.00	.18	.66	.37	-.32	-.25	-.32	-.33	-.28	-.25	-.28	-.27	-.42	.68	-.45
12												1.00	.15	.23	-.22	-.05	-.43	.19	-.05	-.10	-.05	-.21	-.19	.25	-.19
13													1.00	.49	-.49	-.23	-.56	-.22	-.06	-.04	-.02	-.28	-.32	.78	-.30
14														1.00	-.30	-.22	-.56	-.05	-.19	.03	-.18	-.20	-.30	.53	-.41
15															1.00	.00	.41	.23	.07	-.17	.05	.29	.37	-.49	.21
16																1.00	.27	.22	.46	.02	.15	.34	.24	-.26	.45
17																	1.00	.11	.32	.12	.02	.30	.45	-.55	.44
18																		1.00	.32	.25	.00	.29	.22	-.29	.29
19																			1.00	.38	.33	.44	.22	-.19	.53
20																				1.00	.13	.01	-.07	-.06	.14
21																					1.00	.01	.11	-.05	.27
22																						1.00	.42	-.44	.38
23																							1.00	-.41	.44
24																								1.00	-.40
25																									1.00

Table 3

ROTATED FACTOR LOADING FOR TWENTY-FIVE LINGUISTIC VARIABLES

		FACTORS		
		I	II	III
Eigenvalues		9.37	2.51	1.80
Cumulated percentage of eigenvalues		37.5	47.5	54.9
		ROTATED	FACTOR LOADINGS	
1	Mean sentence length (number of words)	0.34	0.57	0.02
2	Mean word length (number of letters)	0.75(1)	0.06	0.06
3	Type token ratio	0.01	-0.00	0.86
4	Word frequency count (absentees from a basic vocabulary list)	0.87	0.06	0.19
5	Suffixed nouns	0.57	0.14	0.20
6	Proper nouns (including adjectivials)	0.22	0.04	0.37
7	Concrete nouns	-0.33	-0.42	-0.07
8	Personal pronouns (3rd person)	-0.27	-0.14	-0.01
9	Personal pronouns (1st and 2nd person)	-0.71	0.24	-0.23
10	Articles and deictics	0.81	0.19	-0.15
11	Ratio of adjectives before nouns to adjectives after nouns	-0.60	-0.16	0.16
12	Adjectives (in the noun phrase employed as attributes)	-0.16	-0.07	0.05
13	Ratio of animate subjects to inanimate subjects	-0.74	-0.03	0.20
14	Short adverbs + adverbials + presentatives	-0.51	-0.27	0.41
15	Prepositions	0.28	-0.01	-0.02
16	Conjunctions	0.15	0.76	-0.18
17	Passives	0.33	0.33	-0.22
18	Comparatives and superlatives	0.13	0.34	-0.04
19	Infinitives	-0.02	0.76	0.21
20	Complement and subject sentences	0.06	0.15	0.19

21	Reciprocals and reflexives	0.02	0.17	-0.02
22	Shift transformations	0.29	0.63	0.33
23	Gerunds and participles	0.22	0.41	-0.04
24	Dialog variable block	-0.77	-0.20	0.08
25	Relative clauses and subordinate sentences	0.08	0.65	-0.18

(1) The level of .6 was selected as the cut-off significance because of the small number of passages and the consequently unreliability of the results. However some of the smaller loadings seem interpratable.

Factor scores were then calculated for each of the 60 passages.

A simple structure does not underlie the 25 variables. Only 54.97 percent of the total variance is accounted for by the first three factors. When attempting ultimately to identify the 22 remaining orthogonal components, it appeared that one variable only exhibited a substantial loading in each of these components.

The first rotated factor exhibits a high positive loading on the following variables (mean word length and word frequency count). Conversely, the same factor exhibited a high negative loading on the following variables (personal pronouns, articles and deictics, ratio of animate subjects to inanimate subjects, dialog variable block). Otherwise when contrasting the passages scoring high in factor I and those scoring low, it clearly appears that factor I can be identified as a content factor (didactic and descriptive reading vs/ light reading).

The second rotated factor exhibited a high loading on following variables : conjunctions, infinitives, shift transformations, relative and subordinate sentences. Using the same methodology as for factor I, one can identify factor II as a syntactical complexity factor.

The third rotated factor exhibited a high loading on one singleton only (type token ratio). This factor can be defined as a redundancy factor.

Instead of a few underlying factors accounting for nearly all of the variance, a large number of singletons emerged from the analysis. The first two common factors accounted for only 47,50 percent of the total variance, with the singleton third factor adding 7.4 percent.

This analysis helps one to understand why nearly all studies which compare materials constructed according to readability formulas in order to observe the effects upon comprehension, fail or lead to inconsistent results (Klare, 1969). When attempting to devise a convenient procedure for making passages easier, this study suggests that one should focuse on many fo the characteristics of written style. Limiting oneself to manipulate the few variables taken into account in the usual readability formulas is deceiving.

Regression on criterion scores

A multiple regression analysis, using factor scores as predictors and passage difficulties measured by cloze as criterion-scores, was made. (Table 4).

Table 4

MULTIPLE REGRESSION ANALYSIS BETWEEN
FACTOR SCORES AND CLOZE RESULTS

	SCHOOL LEVELS		
	5-6	8-9	11-12
Total sum of squares	15.75	9.30	6.79
Mean sum of squares	1.54	1.24	1.59
Percentage of total explained variance	90.22 %	86.67 %	76.58 %
Reduction in sum of squares due to	11.90	6.21	3.47
factor 1 (content factor)	(75.56 %)	(66.77 %)	(51.10 %)
factor 2 (syntactical factor)	1.24	0.60	0.47
	(7.87 %)	(6.45 %)	(6.92 %)
factor 3 (redundancy factor)	0.00	0.41	0.41
	(0.00 %)	(4.41 %)	(6.04 %)

Proportionally, the first factor accounted for more variance at the elementary school level than at the high school level. The syntactical complexity factor accounted for a lower percentage of variance, but this percentage remained about the same across the school levels. The redundancy factor accounted for a very low percentage of variance, but its importance increased across school levels I to III.

SUMMARY

This factor analysis of the linguistic features of written style shows that a large number of factors have to be taken into account to explain the total variance of passage characteristics.

Moreover factors yielded by factor analyses are of varying importance to explain comprehension scores at different ages.

One can hypothsize that the semantic content of the passages is the main difficulty source for youngest children. Other factors (syntactical complexity and redundancy) are, relatively, of no great signifianco so long as the content remains too difficult.

Older children are more and more able to use additional types of cues (such as repetitions of words) for attempting to understand the passages. Moreover, they are more sensitive to syntactical difficulties.

BIBLIOGRAPHY

BRINTON, J.E. and DANIELSON, W.A., A Factor Analysis of Language Elements Affecting Readability, Journalism Quarterly, 1958, 35, pp. 420-426.

BORMUTH, J.R., Development of Readability Analyses, U.S. Department of Health, Education and Welfare, Final Report, 1969, Project 7-0052, Contrat OEC-3-7-070052-0326.

GOUGENHEIM G., RIVENC, P., MICHEA, R. and SAUVAGEOT, A., L'élaboration du français fondamental, Paris : Didier, 1967, pp. 69-113.

HENRY, G., Une technique de mesure de la lisibilité, spécifique de la langue française, Unpublished doctoral dissertation, University of Liège, 1973.

KLARE, G.R., The Measurement of Readability (3rd edition), Iowa City : Iowa State University Press, 1969, pp. 37-80.

STOLUROW, L.M. and NEWMAN, J.R., A Factorial Analysis of Objective Features of Printed Language Presumably Related to Reading Difficulty, Journal of Educational Research, 1959, 52, pp. 243-251.